AFRL-OSR-VA-TR-2013-0061

ENCODING OF 3D STRUCTURE IN THE VISUAL SCENE: A NEW CONCEPTUALIZATION

Christopher W.Tyler, Ph.D., D.Sc, Tai---Sing Lee, Ph.D.

Smith---Kettlewell Eye Research Institute

March 2013

Final Report

DISTRIBUTION A: Approved for public release.

AIR FORCE RESEARCH LABORATORY
AF OFFICE OF SCIENTIFIC RESEARCH (AFOSR)
ARLINGTON, VIRGINIA 22203
AIR FORCE MATERIEL COMMAND

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1. REPORT DATE (*DD-MM-YYYY*)	2. REPORT TYPE	3. DATES COVERED (*From - To*)
29-12-2012	Final	13-09-2009 – 29-09-2012

4. TITLE AND SUBTITLE

Encoding of 3D Structure in the Visual Scene: A New Conceptualization

5a. CONTRACT NUMBER

5b. GRANT NUMBER
FA9550-09-1-0678

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Christopher W. Tyler, Ph.D., D.Sc.,

Tai-Sing Lee, Ph.D. (subcontract)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Smith-Kettlewell Eye Research Institute
2318 Fillmore Street,
San Francisco Ca 94115

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Office of Scientific Research
875 N. Randolph Street, Arlington, VA 22203

10. SPONSOR/MONITOR'S ACRONYM(S)
AFOSR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-OSR-VA-TR-2013-0061

12. DISTRIBUTION / AVAILABILITY STATEMENT

Distribution A: Approved for Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The multidisciplinary goal was to develop an integrated conceptualization of the mid-level encoding of 3D object structure from multiple surface cues. Psychophysical studies showed that depth continuity is a prerequisite for facilitation of Gabor target detection in the context of flanking Gabors, and that, similarly, surface continuity in purely disparity-defined slanted surfaces was strongly enhanced in distributed patch detections as a function of stimulus duration in this dual discrimination task. The unprecedented dips of performance reduction in the component psychometric functions was captured in a computational model based on a novel Leaky Drift Diffusion Theory that we developed for the underlying neural signals, which can serve as an analytic basis for the time course of all neural decision processes. The time course of depth surface perception was studied in a coordinated trio of psychophysical, neurophysiological and functional imaging studies, showing that the perceptual processing of disparity and integration of 3D surface information across depth cues has time courses of several seconds, attesting to complexity of the neural processing hardware. Three complementary computational modeling projects from three collaborating laboratories showed how surface reconstruction could be accomplished across the typically sparse depth information available, and integrated among sparse, incommensurate cue modalities.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE		10	19b. TELEPHONE NUMBER (*include area code*)

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

ACCOMPLISHMENTS

Psychophysical Studies

A simpler structure for local spatial channels as identified with sustained stimuli in the visual periphery. A new evaluation of the local structure of spatial channels with local stimuli in peripheral retina employs the masking sensitivity approach in order to minimize analytic assumptions. The stimuli were designed to assess the range of channel tunings of the predominantly sustained response system in the near periphery. Under these conditions, the range of identifiable channels spans a narrow range of spatial frequencies, from roughly 2 - 8 cy/deg at 2 deg eccentricity to 1 - 4 cy/deg at 8 deg eccentricity. The analysis shows that there are no sustained channels tuned below 2 cy/deg for the central visual field. This two-octave range of channel tuning is much narrower than is conventionally assumed. For local sustained stimuli, human peripheral spatial processing therefore appears to be based on a simpler channel structure than is often supposed.

The role of disparity interactions in the perception of the 3D environment. Disparity interactions across the 3D domain were measured via the masking effect of a Gaussian blob on detection of a Gaussian target was measured as a function of the position, disparity, width and polarity of the mask. This paradigm revealed a large degree of disparity-specific masking that cannot be explained by the masking of its monocular constituents. The masking effects could be modeled as having three additive components, one that has a fixed disparity range and is polarity independent, one with a center/surround form keyed to both the disparity and the polarity of the mask, and one that derives from the monocular masking in each eye. Thus, the profound disparity interaction behavior is not limited to the simple monocular masking properties of the stimuli but reveals extensive connectivity across the disparity domain. Future models of disparity encoding will need to take these properties into account.

Collinear facilitation over space and depth. The detection threshold of a Gabor target can be reduced by the presence of collinear flanking Gabors but is disrupted when the target and the flankers have different disparity. Here, we further investigated whether it is the depth or surface difference between the target and the flanker that causes the abolition of collinear facilitation. The target and the flankers were 1.6 cy/deg vertical Gabor patches with three wavelengths separation between them. There were six viewing conditions: target and flankers were set (A) in the same frontoparallel plane in a collinear configuration; (B) at different disparities but embedded in the same slanted plane, (C) at different disparities in different frontoparallel planes (flankers occupied at the same depth); (D) at different disparities in different frontoparallel planes (flankers occupied at different depth); (E) in the same frontoparallel plane in a non-collinear configuration; (F) at the same disparity but locally slanted. We measured the target contrast detection threshold with and without the flankers present with a temporal 2AFC paradigm with the Ψ staircase method. Strong collinear facilitation was observed when the target and the flankers were either in the same frontoparallel plane or embedded in the same slanted surface even though the target and the flankers were at different disparities. Our results suggest that it is the difference in surface assignment, not the difference in disparity per se, that causes the disruption of collinear facilitation.

Depth spreading through empty space induced by sparse disparity cues. A key goal of visual processing is to develop an understanding of the three-dimensional layout of the objects in our immediate vicinity from the variety of incomplete and noisy depth cues available to the eyes. Binocular disparity is one of the dominant depth cues, but it is often sparse, being definable only at the edges of uniform surface regions, and visually resolvable only where the edges have a horizontal disparity component. In order to understand the full 3D structure of visual objects, our visual system has to perform substantial surface interpolation across unstructured visual space. This interpolation process was studied in an eight-spoke depth spreading configuration corresponding to that used in the Neon Color Spreading Effect. A strong percept of a sharp contour extending through empty space from the disparity edges within the spokes was perceived by all observers. Four hypotheses were developed for form of the depth surface that would be induced by disparity in the spokes defining an incomplete disk in depth: low-level local (isotropic) depth propagation, mid-level linear (anisotropic) depth-contour interpolation or extrapolation, and high-level (anisotropic) figural depth propagation of a disk figure in depth. Data for both perceived depth and position of the perceived contour clearly rejected the first three hypotheses and were consistent with the high-level figural hypothesis for the anisotropic depth propagation for both uniform disparity and slanted disk configurations. We conclude that depth spreading through empty visual space is an accurately quantifiable perceptual process that propagates depth contours anisotropically along their length and is governed by high-level figural properties of 3D object structure.

Hysteresis in Stereoscopic Surface Interpolation. One of the most fascinating phenomena in stereopsis is the profound stereohysteresis in which the depth percept with increasing disparity persisted long past the point of depth recovery with decreasing disparity. To control retinal disparity without vergence eye movements, they stabilized the stimuli on the retinas with an eye tracker. We now report that stereo hysteresis observed by rotating the binocular stereogram image, shows a popout effect as though the depth was rapidly switching on and off, despite the inherently sinusoidal change in the horizontal disparity vector. This stimulus was set up electronically in a circular format so that the random-dot field could be dynamically replaced, eliminating any cue to cyclorotation. Noise density was proportional eccentricity to fade the stimulus near the zero-disparity fixation target, allowing us to verify that fixation was held accurately at zero disparity. For both the invariant and the dynamic noise, profound hysteresis of many seconds delay was found in eight observers for both the onset and offset of the perceived depth surface. A similar hysteresis was obtained for depth popout from vertical disparity modulation of a fixed horizontal disparity. Conversely, sinusoidal modulation of the horizontal disparity to match the horizontal vector component of the disparity rotation did not show the popout effect, which thus seems to be a function of the interaction between horizontal and vertical disparities and is attributable to the time course of surface interpolation processes for the perceived depth structure.

Paradoxical perception of surfaces in the Shepard tabletop illusion. The Shepard tabletop illusion, consisting of different perspective embeddings of two identical parallelograms as tabletops, affords a profound insight in to the nature of the visual processing of surfaces on the basis is of a striking difference in the perceived shapes of the tabletop surfaces. My analysis reveals three further paradoxical aspects of this illusion, in addition to its susceptibility to the ‘inverse perspective illusion’ of the implied orthographic perspective of the table images. These novel aspects of the illusion are: a paradoxical slant of the tabletops, a paradoxical lack of perceived depth, and a paradoxical distortion of the length of the rear legs. The construction of the illusion resembles scenes found in ancient Chinese scroll paintings, and an analysis of the source of the third effect shows that the interpretation in terms of surfaces can account for the difference in treatment of the filled-in versus open forms in Chinese painting from more than 1000 years ago.

Neurophysiological and Functional Imaging Studies

Binocular disparity and relative luminance preferences are correlated in macaque V1, matching natural scene statistics. Humans excel at inferring information about 3D scenes from their 2D images projected on the retinas, using a wide range of depth cues. One example of such inference is the tendency for observers to perceive lighter image regions as closer. This psychophysical behavior could have an ecological basis because nearer regions tend to be lighter in natural 3D scenes. Here, we show that an analogous association exists between the relative luminance and binocular disparity preferences of neurons in macaque primary visual cortex. The joint coding of relative luminance and binocular disparity at the neuronal population level may be an integral part of the neural mechanisms for perceptual inference of depth from images

Recurrent connectivity can account for the dynamics of disparity processing in V1. Disparity tuning measured in the primary visual cortex (V1) is described well by the disparity energy model, but not all aspects of disparity tuning are fully explained by this classic model. Such deviations from the disparity energy model provide us with insight into how network interactions may play a role in disparity processing and help to solve the stereo correspondence problem. Here, we propose a neuronal circuit model with recurrent connections that provides a simple account of the observed deviations. The model is based on recurrent connections inferred from neurophysiological observations on spike timing correlations, and is in good accord with existing data on disparity tuning dynamics. We further performed two additional experiments to test predictions of the model. First, we increased the size of stimuli to drive more neurons and provide a stronger recurrent input. Our model predicted sharper disparity tuning for larger stimuli. Second, we displayed anti-correlated stereograms, where dots of opposite luminance polarity are matched between the left- and right-eye images and result in inverted disparity tuning in the disparity energy model. In this case, our model predicted reduced sharpening and strength of inverted disparity tuning. For both experiments, the dynamics of disparity tuning observed from the neurophysiological recordings in macaque V1 matched model simulation predictions. Overall, the results of this study support the notion that, while the disparity energy model provides a primary account of disparity tuning in V1 neurons, neural disparity processing in V1 neurons is refined by recurrent interactions among elements in the neural circuit.

Neuronal interactions and their role in solving the stereo correspondence problem. Several different approaches can be taken to solving the stereo correspondence problem and there are neurophysiological data that support each of these approaches. There is a particular case for taking advantage of spatial priors to infer disparity because the strategy is so effective, but this method is not mutually exclusive from the other methods discussed. Most of the present neurophysiological data are too general to make strong conclusions about exactly what mechanism or mechanisms the brain might be using. For example, evidence of fast suppression between neurons with very different disparity tuning but with overlapping receptive fields might reflect the stereo uniqueness constraint.

Another general result is that the disparity tuning sharpens over time. Disparity tuning clearly deviates from a Gabor function with narrow peaks and broad valleys, which suggests that the primary visual cortex is refining local disparity estimates over time, as is true of all of the proposed solutions to the stereo correspondence problem that we described.

Visual surface encoding: a neuroanalytic approach. The predominant mode of spatial processing is through a self-organizing surface representation (or attentional shroud) within a full 3D spatial metric. It is not until such a surface representation is developed that the perceptual system seems to be able to localize the components of the scene. The concept of the attentional shroud is a flexible network for the internal representation of the external object structure. In this concept, the attentional shroud is, itself, the perceptual coordinate frame. It organizes ("shrink-wraps") itself to optimize the spatial interpretation implied by the complex of binocular and monocular depth cues derived from the retinal images. It is not until this depth reconstruction process is complete that the coordinate locations can be assigned to the external scene. Spatial form, usually seen as a predominantly 2D property that can be rotated into the third dimension, becomes a primary 3D concept of which the 2D projection is a derivative feature.

The net result of this analysis is to offer a novel insight into the nature of the binding problem. The separate stimulus properties and local features are bound into a coherent object by the "glue" of the global 3D surface representation. Such active binding processes are readily implementable computationally with plausible neural components that could reside in a locus of 3D reconstruction in the human cortex. This locus has been identified by functional imaging as a region of cortex in the dorsal extreme of the lateral occipital complex, adjacent to area V3B. Other aspects of 3D representation were identified as encoding specifically to motion in depth, located more ventrally, anterior to the motion area hMT+ (which encodes not only 2D motion but to frontoparallel motion of 3D cyclopean depth structure defined solely by binocular disparity, in a form that would be invisible to standard motion detectors.)

Theoretical and Computational Analyses

Scene statistics and three-dimensional surface perception. Statistical methods of inference are of great benefit for the analysis of visual perception. By developing and applying efficient statistical inference techniques that consider distributions over 3D shapes, we were able to advance the state of shape from shading considerably. The efficient belief propagation techniques we have developed have similar applications in a variety of perceptual inference tasks. These and other statistical inference techniques promise to significantly advance the state of the art in computer vision and to improve our understanding of perceptual inference in general.

In addition to improved performance, the factor graph approach to shape from shading offers a new degree of flexibility that should allow shading to be exploited in more general and realistic scenarios. Previous approaches have typically relied heavily on the exact nature of the Lambertian reflectance equations, and so could only be applied to surfaces with specific (i.e. matte) reflectance qualities with no surface markings and specific lighting conditions. The factor graph approach applies directly to a statistical model of the relationship between shape and shading, and so does not depend on the exact nature of the Lambertian equation or specific lighting arrangements. Also, the efficient higher-order belief propagation techniques described here make it possible to exploit stronger, non-pairwise models of the prior probability of 3D shapes. However, because the problem of depth inference is so highly under-constrained, and natural images admit large numbers of plausible 3D interpretations, it is crucial to utilize an accurate model of the prior probability of 3D surface. Knowing what 3D shapes commonly occur in nature, and what shapes are a priori unlikely or odd is a very important constraint for depth inference. The factor graph representation of the shape from shading problem can be generalized naturally to exploit other depth cues, such as occlusion contours, texture, perspective, or the da Vinci correlation and shadow cues. The state of the art approaches to the inference of depth from binocular stereo pairs typically employ belief propagation over a Markov random field. These approaches can be combined with our shape from shading framework in a fairly straightforward way, allowing both shading and stereo cues to be simultaneously utilized in statistically optimal way. Statistical approaches to depth inference make it possible to work towards a more unified and robust depth inference framework, which is likely to become a major area of future vision research.

The role of mid-level surface representation in 3D object encoding. Spatial vision is an active process of object representation, in which a self-organization net of neural representation can reach through the array of local depth cues to form an integrated surface representation of the object structure in the physical world being viewed. This conceptualization identifies an active neural coding process that goes far beyond the atomistic concept of local contour or disparity detectors across the field and that can account for some of the dynamic processes of our visual experience of the surface structure of the scene before us. Once the 3D surface structure is encoded, the nature of the elements in the scene can be segmented into the function units that we know as 'objects'. Importantly, this description is compatible with a realization in the neural networks of the parieto-occipital cortex rather than just an abstract cognitive schema.

To realize this schema, visual perception can be viewed as an interpretation of sensory data based on an intrinsic geometry determined by its underlying principles of organization. The perceptual unity relate to the concept of intrinsic constancy under a non-Euclidean geometry, and may be extended to visual modalities such as form, motion, color, depth, etc. The perceptual structure of the visual process can then be described as a topological fiber bundle, with visual space as the base manifold, the mapping from the world to the cortex as the base connection, the motion system as the tangent fiber, and all other relevant visual modalities as general fibers within the fiber bundle. The cross-section of the fiber bundle is the information from the visual scene, an intrinsically invariant (parallel) portion of which represents a visual object. This concept can account for the unity of perceptual binding of the variety of different perceptual cues that are segregated early in the visual process.

Hybrid generative-discriminative classification using posterior divergence. Integrating generative models and discriminative models in a hybrid scheme has shown some success in recognition tasks. In such scheme, generative models are used to derive feature maps for outputting a set of fixed length features that are used by discriminative models to perform classification. In this paper, we present a method, called posterior divergence, to derive feature maps from the log likelihood function implied in the incremental expectation-maximization algorithm. These feature maps evaluate a sample in three complementary measures: (1) how much the sample affects the model; (2) how well the sample fits the model; (3) how uncertain the fit is. We prove that the linear classification error rate using the outputs of the de- rived feature maps is at least as low as that of plug-in estimation. We present efficient algorithms for computing these feature maps for semi-supervised learning and supervised learning. We evaluate the proposed method on three typical applications, i.e. scene recognition, face and non-face classification and protein sequence analysis, and demonstrate improvements over related methods.

Perceptual coding for 3D reconstruction. A primary issue in 3D reconstruction is the real-time efficacy of different coding methods for the multiple decisions among competing 3D solutions. A popular model framework making such coding decisions is the boundary limited drift-diffusion model, which has been developed in parallel in various branches of science from quantum physics to economics. A common property of all such models is the linear increase in variance of the diffusion processes over time, implying an inability to focus on the current information in the environment, and the inevitability of a forced random decision in the absence of any reliable evidence. We have developed an alternative, more plausible model framework for Bayesian information accumulation that solves both problems and provides an accurate account of many features of context effects in human 3D reconstruction performance.

Non-commutative field theory in the visual cortex. The natural mapping from a world space to its phase space is performed by the Bargmann transform obtained via convolution with the coherent states. Then it will be shown that the action of the simple cells, which is exactly a convolution with the receptive filters, performs such a transform. The norm of the output of the simple cells is generally interpreted as an energy function, always positive, output of the complex cells. Hence, the norm of the Bargmann transform, suitably normalized, will be considered as a probability measure. Consequently, to the image it is associated a natural operator to account for probability distribution, that is, the density operator. Consideration of the cortical signal as a Markov process leads, in turn, to a Fokker-Planck equation in the cortical phase space. Its solution expresses the probability that a point with a specific direction belongs to a contour, and it is implemented by the horizontal connectivity in the three-dimensional (3D) cortical space. The output of the Bargmann transform containing information about image boundaries is propagated by the Fokker-Planck equation, resulting in boundary completion and the filling in of the figure.

Ricci flow for 3D shape analysis. Ricci flow is a powerful curvature flow method, which is invariant to rigid motion, scaling, isometric, and conformal deformations. We present the first application of surface Ricci flow in computer vision. Previous methods based on conformal geometry, which only handle 3D shapes with simple topology, are subsumed by the Ricci flow-based method, which handles surfaces with arbitrary topology. We present a general framework for the computation of Ricci flow, which can design any Riemannian metric by user-defined curvature. The solution to Ricci flow is unique and robust to noise. We provide implementation details for Ricci flow on discrete surfaces of either Euclidean or hyperbolic background geometry. Our Ricci flow-based method can convert all 3D problems into 2D domains and offers a general framework for 3D shape analysis. We demonstrate the applicability of this intrinsic shape representation through standard shape analysis problems, such as 3D shape matching and registration, and shape indexing. Surfaces with large non-rigid anisotropic deformations can be registered using Ricci flow with constraints of feature points and curves. We show how conformal equivalence can be used to index shapes in a 3D surface shape space with the use of Teichmuller space coordinates. Experimental results are shown on 3D face data sets with large expression deformations and on dynamic heart data.

3D surface representation using Ricci flow. 3D surface representation has the fundamental importance in middle level vision. This work proposes a rigorous and general framework for representing 3D surfaces with arbitrary topologies in three-dimensional Euclidean space. The geometric information of a surface is decomposed to four layers, topology, conformal structure, Riemannian metric and mean curvature, which together determines the surface uniquely (up to a rigid motion).

Ricci curvature flow is the essential tool to compute the surface representation. Surface Ricci flow deforms the Riemannian metric according to the Gaussian curvature, such that the curvature evolves like a heat diffusion process. Eventually, the curvature is constant everywhere, the final metric represents the conformal structure of the surface. The distortion between the original metric and the final metric represents the original Riemannian metric. This work presents discrete curvature flow methods that are recently introduced into the engineering fields: the discrete Ricci flow and discrete Yamabe flow for surfaces.

Archival publications during reporting period (18 publications):

Chen CM, Chen CC, Tyler CW. Depth structure from asymmetric shading supports face discrimination. *J Vision* (in press).

Huang PC, Chen CC, Tyler CW (2012) Collinear facilitation over space and depth. *J Vision* 12. pii: 20. doi: 10.1167/12.2.20.

Kontsevich LL, Tyler CW (2012) A simpler structure for local spatial channels as identified with sustained stimuli in the visual periphery. *J Vision* (in press).

Li X, Huang AE, Altschuler EL, Tyler CW. Depth spreading through empty space induced by sparse disparity cues. J Vision (submitted).

Li X, Lee TS, Liu Y (2011) Hybrid generative-discriminative classification using posterior divergence. *IEEE Conference in Computer Vision and Pattern Recognition (CVPR)*. 2713-2720.

Potetz BR, Lee TS (2011) Scene statistics and three-dimensional surface perception. In Tyler CW, Ed. , *Computational Vision: From Surfaces to Objects*. Chapman and Hall/CRC: Boca Raton, FL

Samonds JM, Potetz BR, Tyler CW, Lee TS. Recurrent connectivity can account for the dynamics of disparity processing in V1. *J Neuroscience* (in press)

Samonds JM, Potetz BR, Lee TS (2012) Binocular disparity and relative luminance preferences are correlated in macaque V1, matching natural scene statistics. *Proc Nat Acad Sci U S A.* 109:6313-8.

Samonds JM, Lee TS (2011) Neuronal interactions and their role in solving the stereo correspondence problem. In *Vision in 3D Environments*, Ed. Laurence Harris, Michael Jenkin, Cambridge University Press, Ch 7.

Sarti A, Citti G (2011) Non-commutative field theory in the visual cortex. In Tyler CW, Ed, *Computational Vision: From Surfaces to Objects*. Chapman and Hall/CRC: Boca Raton, FL.

Tyler CW, Gill N, Likova LT, Nicholas SC. Hysteresis in stereoscopic surface interpolation. *iPerception* (submitted).

Tyler CW (2011) Paradoxical perception of surfaces in the Shepard tabletop illusion. *i-Perception* 2, 137–141.

Tyler CW (2011) The role of disparity interactions in the perception of the 3D environment. In *Vision in 3D Environments*, L. Harris and M. Jenkins, Eds. Cambridge University Press , Ch 5.

Tyler CW (2011) The role of mid-level surface representation in 3D Object Encoding. In Tyler CW, Ed, *Computational Vision: From Surfaces to Objects*. Chapman and Hall/CRC: Boca Raton, FL

Tyler CW, Likova LT (2011) Visual surface encoding: a neuroanalytic approach. In Tyler CW, Ed, *Computational Vision: From Surfaces to Objects*. Chapman and Hall/CRC: Boca Raton, FL.

Tyler CW, Nicholas SC (2011) Perceptual coding for 3D reconstruction. *IEEE Express* 10.1109/EuVIP.2011.604553

Zeng W, Samaras D, Gu DX (2010) Ricci flow for 3D shape analysis. *IEEE Trans Pattern Analysis Machine Intell (PAMI*), 32:662-677; doi:10.1109/TPAMI.2009.201

Zeng W, Lu F, Yau ST, Gu DX (2011) 3D surface representation using Ricci flow. In Tyler CW, Ed, *Computational Vision: From Surfaces to Objects*. Chapman and Hall/CRC: Boca Raton, FL.

Include any new discoveries, inventions, or patent disclosures during this reporting period (if none, report none):

None